INDUSTRIAL PHD IN BIG DATA AND ARTIFICIAL INTELLIGENCE (XXXVIII ED.)

Curricula "Big data management per la transizione digitale" Università delle Camere di Commercio Italiane "UNIVERSITAS MERCATORUM"

PHD STUDENT: VITTORIO STILE
TUROR: PROF. ROBERTO CALDELLI
RESEARCH PROJECT: RECOGNITION OF AI-GENERATED DEEPFAKES

Guide to Downloading the FaceForensics++ Dataset and Setting Up the Environment

1. Introduction

FaceForensics++ is a large-scale dataset widely used for detecting facial manipulations in videos. It has been created to support research in deepfake detection, offering a variety of video manipulations generated by state-of-the-art techniques. The dataset is available in several compression levels to accommodate different research needs.

2. Requirements and Dependencies

2.1 Frameworks

To work with the FaceForensics++ dataset, the following software frameworks and tools are required:

- *Python*: version 3.8 or later (e.g. reference for this guide Python 3.11).
- *Pip*: a package installer for Python to manage dependencies.
- *Git*: a version control system for cloning repositories.

2.2 Python Packages

Ensure the following Python packages are installed:

- `numpy`
- `scipy`
- `pandas`
- `tensorflow` (required to run predefined models)
- `keras`
- `opencv-python` (for image processing)
- `requests` (for downloading datasets)

2.3 Installation Commands

Use the following commands to install the necessary tools and dependencies:

```
sudo apt-get update sudo apt-get install python3.11 python3-
pip git pip install numpy scipy pandas tensorflow keras
opencv-python requests
```

3. Cloning the Repository

To download the FaceForensics++ dataset, first clone the official GitHub repository using the command below:

```
git clone https://github.com/ondyari/FaceForensics
cd FaceForensics
```

4. Downloading the Dataset

The FaceForensics++ repository provides a script[1] for downloading the dataset with various compression levels and subsets. Below are the details on how to use the script effectively.

[1] Which can be downloaded at the following link: http://kaldir.vc.in.tum.de/faceforensics_download_v4.py

4.1 Basic Download Command

To download the entire dataset, you can use the following command:

```
python download.py —all
```

4.2 Command Parameters

The script offers several parameters to customise the download process:

positional arguments:

output_path	Output directory

Options:

-h	—help		show this help message and exit
-d	—dataset	original_youtube_videos,original_youtube_videos_info,original,DeepFakeDetection_original,Deepfakes,DeepFakeDetection,Face2Face,FaceShifter,FaceSwap,NeuralTextures,all	Which dataset to download, either pristine or manipulated data or the downloaded youtube videos. (default: all)
-c	—compression	raw,c23,c40	Which compression degree. All videos have been generated with h264 with a varying codec. Raw (c0) videos are lossless compressed. (default: raw)
-t	-type	videos,masks,models	Which file type, i.e. videos, masks, for our manipulation methods, models, for Deepfakes. (default: videos)

-n	--num_videos		Select a number of videos number to download if you don't want to download the full dataset. (default: None)
	--server	EU,EU2,CA	Server to download the data from. If you encounter a slow download speed, consider changing the server. (default: EU)

4.3 Example Command

To download the manipulated videos in the “Downloads” folder with light compression (c23) trough the EU2 server, use:

```
python download-FaceForensics++.py /Downloads -d all -c c23
--server EU2
```

5. Dataset Structure and Characteristics

The FaceForensics++ dataset is structured across multiple compression levels, each with distinct file sizes:

- Raw (Uncompressed): ~X.XX GB
- Compressed c23 (Light Compression): ~35,15 GB
- Compressed c40 (Heavy Compression): ~4,84 GB

Each level includes approximately 1,000 original YouTube videos and their corresponding manipulated versions. The manipulations are performed using four different techniques: DeepFakes, Face2Face, FaceSwap, and NeuralTextures.

6. Final Considerations

6.1 Performance

For most use cases, the c23 compression level is recommended as it provides a good balance between quality and file size.

6.2 Resources

Ensure that you have sufficient disk space and a stable internet connection before beginning the download process.

6.3 Support

For any issues or further assistance, refer to the documentation provided in the repository or seek help through my email vittoriostile@gmail.com or LinkedIn profile.